UNCLASSIFIED

AD NUMBER

ADB006851

LIMITATION CHANGES

TO:

Approved for public release; distribution is unlimited.

FROM:

Distribution authorized to U.S. Gov't. agencies only; Test and Evaluation; JUL 1975. Other requests shall be referred to Rome Air Development Center, Attn: IRAP, Griffiss AFB, NY 13441.

AUTHORITY

RADC ltr, 12 Apr 1978

THIS PAGE IS UNCLASSIFIED

THIS REPORT HAS BEEN DELIMITED
AND CLEARED FOR PUBLIC RELEASE
UNDER DOD DIRECTIVE 5200.20 AND
NO RESTRICTIONS ARE IMPOSED UPON
ITS USE AND DISCLOSURE.

DISTRIBUTION STATEMENT A

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

AD B006851

RADC-75-188
Final Technical Report
July 1975

VOICE INPUT CODE IDENTIFIER

Threshold Technology Inc.

DDC
RECEIVED
OCT 6 1975
C

Distribution limited to U. S. Gov't agencies only; test and evaluation; July 1975. Other requests for this document must be referred to RADC (IRAP), Griffiss AFB NY 13441.

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York 13441

This technical report has been reviewed and approved for publication.

APPROVED: Richard S. Vonusa

RICHARD S. VONUSA
Project Engineer

APPROVED: Howard Davis

HOWARD DAVIS
Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER: Carlo P. Crocetti

CARLO P. CROCETTI
Chief, Plans Office

Do not return this copy. Retain or destroy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-75-188	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* VOICE INPUT CODE IDENTIFIER		5. TYPE OF REPORT & PERIOD COVERED Final Technical Report
		6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) Phillips B. Scott		8. CONTRACT OR GRANT NUMBER(s) F30602-74-C-0171
9. PERFORMING ORGANIZATION NAME AND ADDRESS Threshold Technology Inc. Route 130 & Union Landing Road Cinnaminson NJ 08077		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 31011F 40270508
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRAP) Griffiss AFB NY 13441		12. REPORT DATE July 1975
		13. NUMBER OF PAGES 45
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT *(of this Report)*
Distribution limited to U. S. Gov't agencies only; test and evaluation; July 1975. Other requests for this document must be referred to RADC (IRAP). Griffiss AFB NY 13441.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*
Same

DDC
RECEIVED
OCT 6 1975
C

18. SUPPLEMENTARY NOTES
RADC Project Engineer:
Richard Vonusa (IRAP)

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*
Speech Recognition
Pattern Recognition
Acoustic Phonetics

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*
This report describes the development, operation and performance characteristics of an Advanced Development Model of a Voice Input Code Identifier (VICI). The VICI is an isolated word recognition system capable of recognizing the English digits and four control words, CANCEL, ERASE, VERIFY and TERMINATE. The system will accept these words independent of speaker for a large population of General American males. No training of the system by a speaker is necessary. By the use of an alphanumeric output display, a speaker using the system can

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

verify that each digit spoken into the system was correctly recognized. Errors can be corrected through the use of the control words.

To confirm system performance several final tests were held, two of which included live inputs rather than tape recordings. Individual digit recognition accuracy in each of two tests from magnetic tape was 98.7 percent for a total of 65 speakers. In the live tests a total of 30 speakers each spoke into the system's 75 groups of digits, each group consisting of four digits followed by the word VERIFY to simulate operational conditions. Individual digit accuracy in these tests was 97.9 percent for 30 speakers. Approximately 92.5 percent of all digit groups were inputted and verified without error. The remaining groups were corrected and properly entered. With feedback verification and error correction all talkers were able to enter all digit groups correctly. Most codes, together with the verify command, were entered in four to seven seconds when no errors were detected. Typically, 10 to 12 seconds were required to observe and correct a digit error and enter the corrected code.

The VICI system is based upon the VIP-100 isolated word recognition system which normally requires the input of training data by each talker who uses the system. For use in the VICI application both hardware and software modifications were made to a VIP-100 system to allow recognition of the VICI vocabulary spoken by a large speaker population without adaptation or training by any speaker from a large population of General American males.

TABLE OF CONTENTS

Section		Page
I	BACKGROUND AND INTRODUCTION	1
II	TECHNICAL DISCUSSION	4
	A. Introduction	4
	B. Basic Approaches to Automatic Speech Recognition	4
	C. Description of the VIP-100	6
	1. Preprocessor	6
	2. Feature Extractor	8
	3. Minicomputer Function	8
	4. Training	8
	5. Recognition Mode of Operation	9
	D. Development of a Universal Reference Array Set	9
	1. Alternate Reference Arrays	9
	2. Merging Reference Arrays	12
	E. One Word Training Sample Experiments	19
	F. Recognition Networks for Universal Speaker Sets	20
	G. VICI Software	29
III	FINAL SYSTEM TESTS	35
	A. Background of Test Data	35
	B. Final Testing from Tape	35
	1. Random Digit and Control Word Test Results	38
	2. Four-Digit Group Tests	38
	C. Final Testing with Live Inputs	38
IV	CONCLUSIONS AND RECOMMENDATIONS	43
	A. Conclusions	43
	B. Recommendations	43

LIST OF ILLUSTRATIONS

Figure		Page
1	VIP-100 automatic speech recognition system	3
2	Pattern recognition process	4
3	Block diagram of VIP-100 speech recognition system	7
4	Correlation scores for 10 talkers for digit "zero"	11
5	Reference array for word ERASE resulting from merging training data for talkers MH and EC. Encircled points are not common to both talkers	13
6	Reference arrays for two talkers for word ERASE	14
7	Correlation score matrix resulting from correlation of training data representing each word of each of 15 speakers with word 3 (the digit "three") of speaker 1	17
8	Composition of merged training samples used for five-sample merge	18
9	Error matrix for 50 speakers speaking 50 groups of four digits each (No Training)	23
10	Error matrix for 50 speakers speaking 50 groups of four digits each with single repetition training on the digits 1, 3, and 9	24
11	Logic diagram and equivalent logic equation for /ʃ/ recognition net word	30
12	Error matrix for 34 speakers reading VICI vocabulary list without second-look	33
13	Error matrix for 34 speakers reading VICI vocabulary list with second-look	34
14	Measured near-field frequency response of Telex Model 1200 microphone used for making VICI data tape recording. Microphone measured at 1/4" distance from output orifice of a calibrated Plane-Wave-Tube	36
15	Error matrix of 45 speakers each uttering 280 digits and control words in a random arrangement	39

LIST OF TABLES

Table		Page
I	Test Data Words in Random Order	15
II	List of 50 Four-Digit Groups Used for Tests With Training Digits	21
III	Results of Single Digit Train Experiments	22
IV	Phoneme-Like Feature Recognition Logic Equations	25-28
V	List of 75 Four-Digit Groups Used for Final Live Tests at TTI and at RADC	37
VI	Results of Live Test Held at TTI with 10 Speakers Each Inputting 75 Four-Digit Groups	40
VII	Results of Live Test Held at RADC with 21 Speakers Each Speaking 75 Four-Digit Groups	41

EVALUATION

This report represents a major achievement in the area of automatic speech processing. It proved that it is possible to achieve high word recognition scores in real time using a limited vocabulary with words spoken in a discrete manner and independent of speaker for male speakers regardless of geographic accent. With the aid of visual feedback, all errors were able to be corrected thus insuring proper data entry into the machine.

Because of the success of this program, many practical applications are now emerging. For example, the Voice Input Code Identifier (VICI) will be used in conjunction with the ESD Base and Installations Security System's "Automatic Speaker Verification" (ASV) system. The ASV system which was developed by RADC uses the voice characteristics of an individual as a means of authenticating him for entry control. Presently, the ASV system requires the individual to identify himself with a four digit code by an input device such as a keyboard or badgereader. VICI shall eliminate the need for input devices and will allow an individual to "speak" his code numbers as a means of identifying himself to the verification system.

In addition, this word recognition technology will be transitioned into a natural USAF application. A voice actuated system shall supply pertinent information to a computer as an aid for cartographers. Present mapping techniques require a cartographer to position a X - Y reader device over a smooth sheet, read the required bathometric numbers via the map and then enter these digits to a computer which correlates them with the positioning device. This process of turning away from the table to enter numbers via the manual keyboard diverts the operator's attention and tends to slow down the data entry process. By utilizing a word recognizer, the operator can speak the required digits and enter them automatically into the computer without losing sight of the manuscript. The voice system is more efficient in that it will reduce the data entry time which presently averages 12 seconds to an average of 3 seconds.

These applications and others will insure that voice-controlled devices will have a valuable role in future information processing systems.

Richard S. Vonusa

RICHARD S. VONUSA
Project Engineer

Section I

BACKGROUND AND INTRODUCTION

The application of an automatic speech recognition (ASR) system as a front-end for the Base and Installation Security System's (BISS) automatic speaker verification system can provide a more reliable means of entering speaker verification data. An automatic speaker verification experimental model was fabricated under RADC Contract F30602-72-C-0294. To use this verification system it was first necessary for an individual to manually enter, via a keyboard, a sequence of digits to alert the system as to his identity. This manual data entry can now be eliminated by the use of "Voice Input Code Identifier" (VICI) system which has been developed during the contract described in this report. The combination of the VICI system and the speaker verification system can provide implementation of a fully automatic voice oriented technique to allow an individual requesting base entry to claim identity and be verified. Thus, the need for picture badges, the keypunching of code numbers and other fallible mechanical methods of an individual claiming his valid identity will be eliminated.

The VICI system has been developed to recognize with very high accuracy the English digits zero through nine, plus the control words CANCEL, ERASE, VERIFY and TERMINATE independent of speaker for a large population of General American males. A feedback system has been incorporated to allow the speaker to verify each digit entry and if necessary to correct a faulty entry by the use of the control word, ERASE, and then enter a new digit. A complete code group of four digits can be accepted by the use of the control word VERIFY or rejected by the word CANCEL. The speaker can view on an alphanumeric display each recognized digit within .1 to .2 seconds after it is pronounced in order to verify the correctness of each digit entry. Live tests involving a total of 30 speakers showed that a four digit group could be entered into the VICI system with verification in as short an interval as 2.8 seconds. Four to seven seconds were typically required for most speakers for a digit group if no errors were made either by the speaker or the system. Ten to 12 seconds were required by most speakers to detect and correct an error and complete the entry of a proper code. It was necessary to employ correction for an average of 7.5% of the 75 digit groups spoken by the 29 participants in the live tests. In <u>every</u> case the errors were correctable and <u>every</u> code was entered properly.

In addition to the live tests which were conducted just prior to and at the time of delivery to RADC of the VICI equipment, several tests series were conducted by the use of magnetic tape recordings of a total of 65 male speakers only 11 of whom were used subsequently in the live tests. The speakers who made the tape recordings over a period of several months ranged in age from 16 years to 65 years. The majority of these speakers were in the 20 to 40 year age bracket. Overall, therefore, the VICI system has been tested by 83 male speakers.

The VICI system developed for this contract is based upon the Threshold Technology Inc. (TTI) commercial VIP-100 limited vocabulary isolated word recognition system. The VIP-100 normally requires training (adaptation) by each talker using it. This training is accomplished by inputting five to 10

samples of each vocabulary word by each user. The VIP-100 which served as the basis for the VICI system was modified in both hardware and software to allow operation without the necessity of entering any training data for each speaker.

The VIP-100 system includes a speech preprocessor and a minicomputer, the Nova 1200 manufactured by Data General which includes 8K of core memory. For verification, a display module based upon a Burroughs Self-Scan alphanumeric display panel is included. The display has a 32 character memory and is capable of displaying 16 characters at a time. The microphone used in the development of the VICI system is a Telex model 1200 which is a noise-cancelling unit. An ASR 33 Teletype has been supplied for control and data input/output functions. Figure 1 is a photograph of a VIP-100 system.

Section II of this report describes the basic approaches to speech recognition which led to development of the VIP-100, together with a description of the operating principles of the VIP-100. Next, the development of universal talker data characteristics is discussed. Experiments with single-repetition training samples are described, followed by a description of the hardware and software modifications necessary to accomodate a large talker set without individual training. A description and the results of final system tests, both live and from tape are included in Section III. Conclusions and recommendations are listed in Section IV.



Figure 1. VIP-100 automatic speech recognition system.

Section II

TECHNICAL DISCUSSION

A. Introduction

In order to best meet the requirements of this program in the development of an Advanced Development Model Voice Input Code Identifier, an existing VIP-100 word-recognition system was modified in both hardware and software for this application. The VIP-100 system was previously developed by TTI for commercial use as a limited-vocabulary isolated-word recognition system with adaptation for each speaker using the system as a requirement for its operation. In the following paragraphs of this section the factors leading to the development of the VIP-100 are presented followed by an outline of the operation of a VIP-100. The investigation leading to the hardware and software modification of the VIP-100 system for use without adaptation is then explained. The VIP-100 system supplied to this program includes a speech preprocessor, a Nova 1200 minicomputer manufactured by Data General Corporation with 8K of core memory, and a 16 character alphanumeric display module. A Telex model 1200 noise cancelling microphone is used for speech input to the system and a Teletype Model ASR 33 is used for control and data input/output functions.

B. Basic Approaches to Automatic Speech Recognition

Four processing functions are common to all automatic speech recognition systems. These functions as shown in Figure 2 consist of a microphone trans-

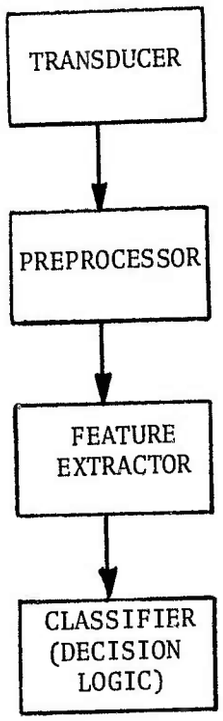


Figure 2. Pattern recognition process.

ducer, a preprocessor, feature extractor and a final decision level classifier. Early attempts at automatic speech recognition either deleted entirely the feature extraction process or utilized a simplified form of template matching. Experience with template matching soon led to the realization of its limitations. Slight variations of the individual speech samples of a particular word would result in gross misclassifications. This limitation resulted in the impractical requirement for a large memory containing a pattern and all its prototypes.

Considerable mathematical formalism has been developed for various automatic speech recognition processes. However, no general theory exists which can preselect the information bearing portions of the speech signal. Therefore, the design of the feature extractor is heuristic and must use ad hoc strategy. Only actual experimental data can determine the value of a particular feature set. It is this particular dilemma which has resulted in the recent increased emphasis given feature extraction research for pattern recognition systems.

It is possible to form many transformations of the speech signal which would enhance certain properties and make them more easily detectable in an automatic speech recognition system. However, speech is neither periodic nor aperiodic, but must be considered as a quasi-periodic signal so that analytical techniques that are developed must reflect temporal features of significance as well as spectral features. Maintaining this dual viewpoint throughout the analysis requires a modification of classical time-domain and frequency-domain analytical techniques. To retain both of these characteristics in a frequency analysis, a method which produces a short-duration spectrum is essential.

Frequency-domain representation of the speech signal is particularly advantageous since (1) it is known that the human auditory system performs a crude frequency analysis at the periphery of auditory sensation and (2) because it has been shown, by acoustical analysis of the vocalization system, that an exact description of the speech sounds can be obtained with a natural frequency concept model of speech production.

A periodic function of time possesses a power spectrum with finite amounts of power located a discrete points in the spectrum, commonly described as a line-spectrum. An aperiodic function that contains finite energy and is Fourier-transformable possesses an energy density spectrum that is a continuous function of frequency. For analyzing speech signals, it is desirable to obtain the spectral energy distribution and its variations as a function of time. Sufficient resolution must be maintained in both the frequency and time domains so that all of the information-bearing properties in both domains can be detected.

Spectrum analysis can be achieved by direct analog circuitry, through the use of the Fast Fourier Transform (FFT) and a high speed digital computer or by the use of linear predictive analysis. In all of these methods, equivalent problems occur. The FFT produces a discrete spectrum which, with a sufficiently high sampling rate, approaches that of the continuous Fourier Transform. Many different types of data windows have been utilized in the

FFT. The choice of the window is similar to the choice of the filter response in the analog spectrum analyzer. A "picket fence" effect can occur both in the FFT and the analog spectrum analyzer representing the contributions of the individual filters in the analog analyzer or the separate coefficients of the various terms in the FFT calculation. Analogous problems are introduced using linear predictive analysis in the selection of the number of coefficients employed in the process. In all cases, however, spectrum analysis is only the first step in the feature extraction process. Considerable additional processing is required in order to achieve the detection and recognition of the information-bearing elements (significant features) of the speech signal which has been transformed to accentuate these elements in the spectrum analysis process.

The final processing level after the recognition of the elemental speech units is the word decision logic. For isolated words, it is possible to examine the phonetic sequences produced by a feature extractor and to determine the closest match to a set of stored reference samples. The decision involving the closest match is made at the end of the word and can be achieved with relatively simple processing techniques. These reference samples can be obtained from a particular talker as in a trainable speech recognition system or can be universal samples as have been developed during this investigation.

The VIP-100 speech recognition system, designed by TTI as a general purpose speech recognizer has served as the basis for the VICI advanced development model developed during this contract. The VIP-100 employs the processing functions just described.

C. Description of the VIP-100

The VIP-100 was originally designed to recognize a vocabulary, essentially unrestricted in content (but restricted in size by the storage limitations of the core memory of the associated minicomputer) with automatic adaptation for individual speakers and words. This system has been modified to allow the recognition of a specific fixed vocabulary by an unlimited speaker set without adaptation to individual speakers. The modifications which were made to the VIP-100 are described in Section II.F.

Operation of the basic VIP-100 is described in the following paragraphs. Figure 3 is a block diagram of the system as originally designed. Both the preprocessor and feature extractor functions are hardwired. The classifier function is performed by software in a Data General Nova 1200 minicomputer. The minicomputer also time normalizes word durations and provides core storage of the reference patterns for each word in the vocabulary.

1. Preprocessor

The initial section of the preprocessor shapes the output from the microphone to remove irregularities and produce a normalized speech spectrum. This equalized signal is then passed through a real-time spectrum analyzer consisting of a bank of 19 contiguous active bandpass filters ranging in center frequency from 260 Hz to 7626 Hz. The outputs of the filters are full-wave rectified and logarithmically compressed. This latter operation provides

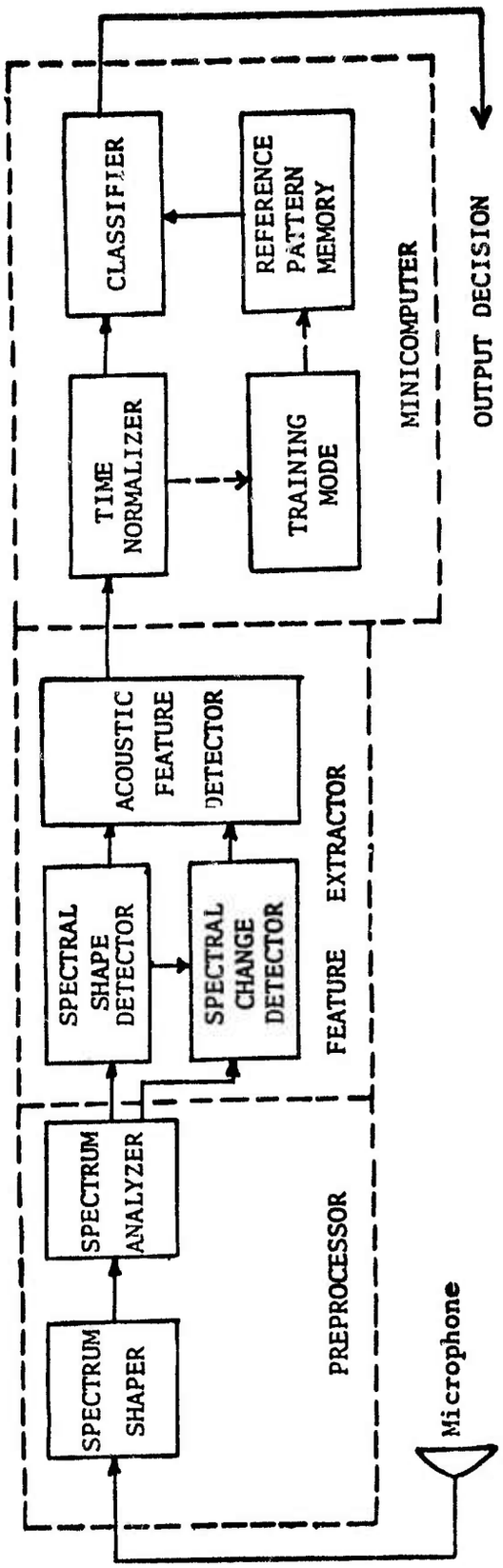


Figure 3 Block diagram of VIP-100 speech recognition system.

a 50dB dynamic range and produces ratio measurements when subsequent features are derived from summation and differencing operations, thereby minimizing the input amplitude dependence.

2. Feature Extractor

The function of the spectral shape detector is to develop spectral derivative (dE/df) features indicating the overall spectrum shape. The spectral shape and its changes with time are continuously measured over the frequency range of interest. Combinations and sequences of these measurements are processed to produce a set of significant acoustic features.

The features used in the VIP-100 are a selected subset (including complex combinations) of 32 acoustic features. Each feature is extracted by a combination of analog operations and binary logic. The output of the feature extractor consists of 32 binary signals, F_1, F_2, $\ldots F_{32}$.

The features are of two types, primary features and phonetic-event features. Features of the former category describe the spectrum directly by indicating local maxima and areas of increasing or decreasing energy with frequency (slopes). The latter category consists of features which represent measurements corresponding to phoneme-like events. Included in this set are vowels, nasals and fricatives.

3. Minicomputer Functions

The minicomputer performs the functions shown in Figure 3. For a spoken word, the 32 encoded features and their time of occurrence are stored in a short term memory. When the end of the utterance is detected by the feature-extractor logic, the duration of the word is divided into 16 time segments and the features are reconstructed into a normalized time base. The pattern-matching logic subsequently compares these feature occurrence patterns to the stored reference patterns for the various vocabulary words and determines the "best fit" for a word decision. 512 bits of information (32 features mapped into 16 time segments) are required to store the feature array of an utterance or reference pattern.

4. Training

The training mode of the operation is a necessary prelude to the normal operation of a VIP-100 system when the system is used as a word recognition system which is adaptable to individual talkers. The VIP-100 which has been modified for VICI use with a universal speaker set does not normally require training (adaptation) by a particular talker. However, the ability to be adapted to or trained for each speaker has been retained in the VICI VIP-100. Furthermore, a series of experiments were conducted involving the use of single training samples for certain digits for increased accuracy. These experiments are described in Section II.E of this report. During the training mode of a conventional VIP-100, or the VICI VIP-100, a time-normalized feature array is extracted for each repetition of a given word. A consistent array of feature occurrences (between repetitions) is required before the features are stored in the reference pattern memory. A template threshold

factor is chosen such that a feature occurrence (in a given time segment) is considered valid only when it occurs a minimum number of times relative to the number of training samples. Usually, this threshold factor is set to be between 30-50% of feature occurrences within the training samples.

5. Recognition Mode of Operation

In the operational mode, each new word spoken into the system is processed in a manner analogous to the training procedure--i.e., feature extracted, digitized and time normalized. The resultant test word array then is compared digitally to the stored reference array for each vocabulary word. Similarities and dissimilarities in each compared array are appropriately weighted and the net result provides a weighted correlation product. Correlation products also are generated after shifting the input word array ± 1 time segment. The stored reference word array producing the highest overall correlation is selected as the test word. This decision is then displayed to the speaker in an appropriate manner for verification of accuracy.

D. Development of a Universal Reference Array Set

As previously mentioned, an important preliminary phase in the operation of the VIP-100 system is adaptation of the system for the voice of a particular user by means of inputting training samples of each word in the vocabulary. During this adaptation or training phase, each vocabulary word is pronounced by the user from one to 10 times. Usually, 10 repetitions of each word in the training phase are used in order to assure maximum recognition accuracy. It has been observed, however, that a single training word for each vocabulary word is adequate for good accuracy if the training words are spoken in close time proximity to the test data inputs. Therefore, a possible mode of operation of the VICI system would be to input the VICI four-digit code, preceded by a complete training phase with a single word training sample. However, such a procedure is undesirable from an operational standpoint because of the time required. Experience has shown that at least one second per word would be necessary during the training and recognition phases with naive speakers who would typically use such a system in the field. Therefore, at least 20 seconds would be required for the training of the 14 word vocabulary of digits plus control words as well as inputting the four digit code number and using one or two additional control words. A realistic limit of 10 seconds for inputting the entire message including any training and verification was established for the system by RADC. Therefore, it became obvious early in the VICI program that the development of prototype reference arrays which are representative of large groups of speakers would be necessary in order to achieve the required recognition accuracy. A minimal training period of five seconds allowed in the specification for the advanced development model could be used for single sample training of two or perhaps three digits which were the most difficult to recognize accurately with universal reference arrays. Experiments with the use of single digit training samples for certain digits will be discussed later.

1. Alternate Reference Arrays

Several different approaches were explored in the attempt to develop an optimum reference array set for universal speaker use. The first of these

approaches involved the use of alternate reference arrays for each vocabulary word chosen such that each array represented a wide variety of expected pronunciations for each word. In many of the commercial applications in which the VIP-100 has been used, it has been noted that a particular talker has been able to achieve highly accurate recognition for a large number of words, especially digits, when using another speaker's stored reference arrays. Often when this phonomenon occurs the two speakers have been found to have been raised in the same geographical area. Therefore, their pronunciation of words is similar and insofar as the ASR system is concerned, they are essentially identical. It should be possible by storing alternate sets of prototype referance arrays for each of the required vocabulary words, to accomodate a large group of talkers from different geographic areas and to achieve good recognition accuracy without additional training or adaptation for any individual speaker using the system.

In order to conduct initial experiments with the use of alternate reference arrays, the 14 word VICI vocabulary was recorded on audio tape by a total of 20 talkers. Each talker repeated each vocabulary word ten times as he would in a normal VIP-100 training phase. A set of special purpose versions of the general VIP-100 training and recognition computer programs were constructed in order to allow data from the VIP-100 preprocessor to be inputted to TTI's real-time disk operating system (RDOS). The use of a disk memory in conjunction with a digital computer provided for the storage of large amounts of training data in a convenient form. This special experimental software also was designed to produce correlation score matrices for a variety of conditions of talkers and word combinations from the data stored on disk. These correlation scores were calculated in the same manner as the correlation products used to choose the proper word in the recognition mode of operation previously described. In the recognition mode, however, the correlation products represented a comparison of reference word arrays stored by the talker using the system in the training phase with the word array resulting from an unknown input word spoken by the speaker who had trained the system. For these experiments, these correlation products resulted from comparisons of the same word as spoken by various talkers, or different words spoken by the same talker, or different words spoken by different talkers. The matrices formed from these correlation products effectively allowed comparisons of pronunciations of various words and illustrated the similarity and dissimilarity of different words. Initially, correlation matrices for each of the 14 words in the VICI vocabulary were constructed. Figure 4 illustrates an abbreviated matrix for 10 talkers for the digit zero. The matrix in the figure shows the correlation scores which were calculated when the training data array for each of the 20 talkers for the word Zero was compared with each of the other 20 talkers for the digit zero. The on-diagonal elements of the matrix are equivalent to a self correlation which is simply two times the total number of points in the 32 x 16 array generated by the particular talker as a consequence of the algorithm which computes correlation products. These matrices were then examined to determine which talkers exhibited the best training data correlation with other talkers for each word and which talkers showed the poorest correlation for that word. Training data for the five talkers representing the best and the worst correlations were chosen for each of the 14 vocabulary words. A reference data set was then established by the use of these choices. Each vocabulary word in the set of 14 then had five alternative reference samples

WORD CORRELATION SCORES

WORD#	SPEAKER#	01	02	03	04	05	06	07	08	09	10
00	01	252	103	120	76	121	158	101	151	146	89
	02	103	302	147	167	132	125	160	142	153	107
	03	120	147	308	144	141	118	145	143	146	134
	04	76	167	144	276	161	111	158	123	130	128
	05	121	132	141	161	298	123	131	160	131	125
	06	158	125	118	111	123	272	135	137	176	79
	07	101	160	145	158	131	135	306	152	159	122
	08	151	142	143	123	160	137	152	310	141	107
	09	146	153	146	130	131	176	159	141	288	94
	10	89	107	134	128	125	79	122	107	94	236

Figure 4. Correlation scores for 10 talkers for digit "zero".

from five different talkers. A test data base was recorded by 19 talkers of the 20 who originally recorded the training data base. This test data base, recorded approximately two weeks after the training data recordings were made, included a total of 20 repetitions of each of the 14 vocabulary words spoken in a random order. Table 1 illustrates the test data set used in this and subsequent tests. The list which contains 140 words was read two times.

2. Merging Reference Arrays

Preliminary tests involving the use of five different speakers for each word disclosed recognition problems with a few talkers on certain words. Most of these recognition problems were associated with talker and word combinations not represented in the training set. Therefore, merging of training data was tested next. Five training samples of each word from each talker not previously included in the training set were merged with five samples of training data from a talker previously included in the training set. This merging of training data was accomplished simply by the use of a conventional VIP-100 software training routine in which the computer was instructed to input ten samples of each word as training data and form a reference matrix from those trained ten samples as is usually done in a commercial VIP-100. Recognition accuracy using these merged multiple representations improved to 96.9 percent correct recognition of 280 words from Table I as spoken by each of 14 talkers, all of whom were also included in the training data set.

Figure 5 illustrates such a reference array (for the word "Erase) which has been generated in this manner from the two talkers, MH and EC. Individual reference arrays for "Erase" for these two speakers are shown in Figure 6. The points in Figure 5 which are encircled were contributed by one or the other but not both of the talkers. All other points appeared in the individual reference arrays for each of the talkers. In a few cases (not shown) points which appeared for one or the other talker were not included in the reference array because they did not meet the threshold criteria established. Since it is possible to increase the number of training samples required to generate a reference matrices it should be possible to merge a multiplicity of talkers speaking the same word in order to obtain an overall average pronunciation for that word. Experiments with the merging of reference arrays from a multiplicity of talkers were conducted at a subsequent time and will be discussed later.

At this point in the program, detailed studies were made of the reference data arrays derived from the original 20-speaker training data. Comparisons between the various talkers revealed that certain features (especially maxima above the second vowel formant) which heretofor had been included among the 32 recognition features varied significantly from speaker to speaker for some words. Therefore, correlation matrices for the 20 talkers speaking 14 words were created using a new feature set. This set consisted of 10 maxima instead of the original 17, plus 6 negative slope features and 16 phoneme and class features. These new matrices were used, as before, for the manual selection of reference data from five talkers for each word. Recognition tests were then conducted for the same 14 talkers as were represented in the previous test. The same test data base was used. Overall accuracy improved from 96.9 percent to 97.2 percent.

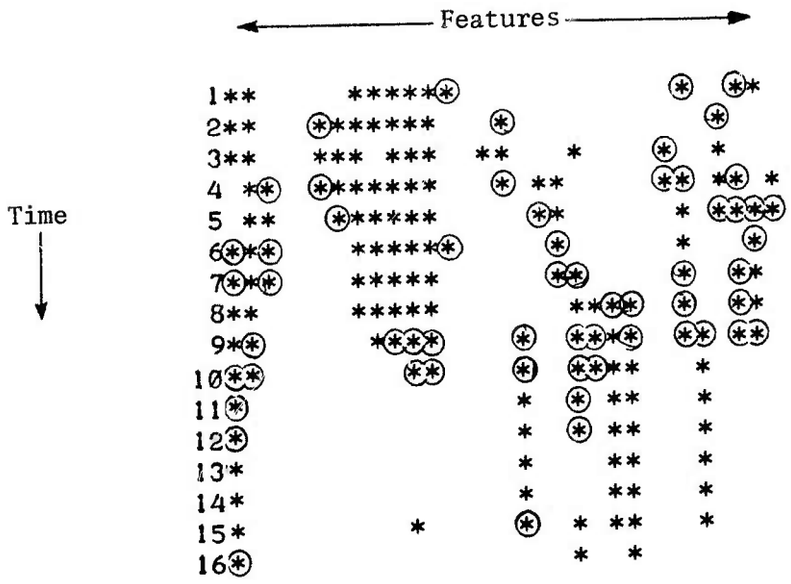
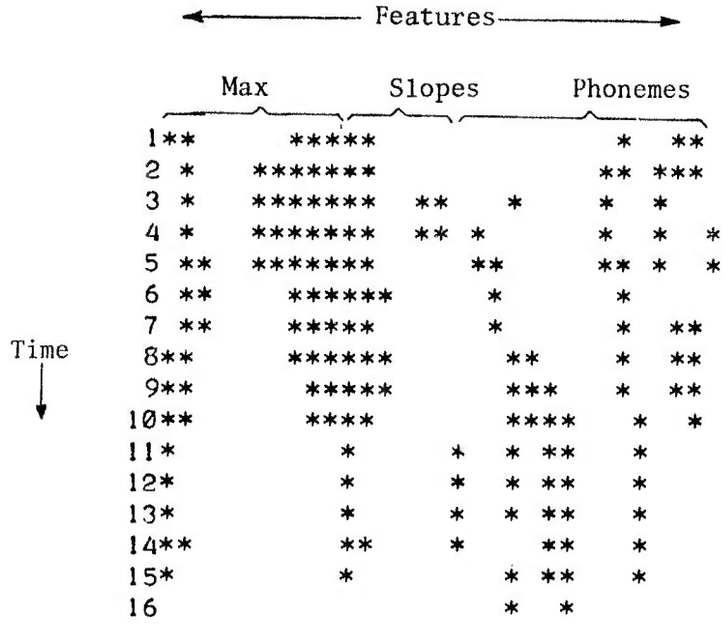
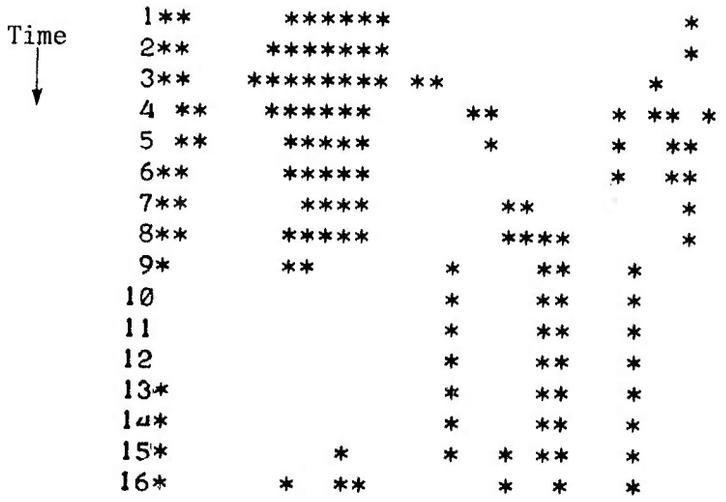


Figure 5. Reference array for word ERASE resulting from merging training data for talkers MH and EC. Encircled points are not common to both talkers.



Talker MH



Talker EC

Figure 6. Reference arrays for two talkers for word ERASE

TABLE I TEST DATA WORDS IN RANDOM ORDER

8	7	VERIFY	3	CANCEL
0	TERMINATE	8	2	ERASE
4	7	3	1	9
5	4	6	5	6
1	0	1	VERIFY	2
ERASE	3	2	0	CANCEL
5	2	9	9	VERIFY
ERASE	ERASE	3	0	8
CANCEL	TERMINATE	9	8	3
5	2	CANCEL	CANCEL	CANCEL
3	0	1	TERMINATE	5
4	TERMINATE	6	8	TERMINATE
VERIFY	0	0	8	4
1	VERIFY	2	6	7
3	CANCEL	VERIFY	4	7
3	0	3	5	1
VERIFY	7	6	4	CANCEL
7	ERASE	9	9	6
ERASE	9	8	0	ERASE
VERIFY	9	1	VERIFY	8
1	TERMINATE	5	2	VERIFY
,	5	7	CANCEL	ERASE
4	2	5	9	6
CANCEL	TERMINATE	4	4	9
TERMINATE	8	3	7	2
6	7	6	TERMINATE	0
1	8	ERASE	5	TERMINATE
1	6	4	2	ERASE

The computer program used for calculating and printing correlation scores described above was expanded to allow calculation of correlation scores between the training data for all words of all talkers of the set stored on a disk memory. With this program modification it was possible to estimate the usefulness of the training data of a particular word in the VICI vocabulary from a particular talker as a possible universal training sample, for a given number of talkers. Figure 7 illustrates a correlation matrix resulting from this program revision. The correlation scores for talkers 01 through 15 for words 00 through 13 compared with word 3 (the digit three) for talker 01 are shown in this figure. Columns in the figure are words, rows are talkers. Note the relatively high scores in the column of word 3 as compared with the remainder of the matrix. As would be expected, the correlation score for the digit "three" for speaker 1 is the highest in the matrix because it is the self-correlation score for this word and speaker. The correlation score for word 3 of speaker 9 is lower than the score for word 8 of speaker 9. This phenomenon indicates that if word 3 from speaker 1 were to be used as training data, word 8 as spoken by speaker 9 would be more likely to be recognized as a "three" than would word 3 from that speaker. Therefore, the use of this sample from speaker 1 is questionable at best. In a few cases such phenomena occurred several times in a particular matrix, rendering that particular sample word unsatisfactory. This procedure does not guarantee that "three" will always be correctly recognized because this matrix does not consider the potential training data for other words. It does, however, provide an effective way to reject possible training data samples which would obviously be unsuitable.

By the use of this "global" correlation technique, all words from the original training data set of 20 talkers were individually correlated against this training data set. Each word sample was then evaluated manually and ranked for suitability as reference data. Another program modification allowed merging of training data stored on disc on a word by word basis. For a particular word, training data from two to 10 talkers could be merged to form a new training sample. The merged data could then be put out on paper tape to be used with the VICI system as reference data.

A 14 word reference array set was constructed by merging data from the five highest ranking speakers for each word in the global correlation of 20 talkers mentioned above. This ranking was based upon examination of correlation matrices such as that shown in Figure 7 for all 14 words of the VICI vocabulary from 20 speakers (280 matrices total). The example shown in Figure 7 would be placed in the questionable category. Any other instances of a score which is higher for a word other than the one for which the matrix was generated put that word sample in the "bad" category. Conversely, significantly higher correlation scores for the same word as generated by other speakers, as compared with other words, resulted in a ranking of good or very good for a training sample candidate. Figure 8 illustrates the speakers whose training data were merged for each word in the VICI vocabulary to form this new training data set. At least one sample was taken from each speaker with as high as eight samples from certain speakers. A test of 29 talkers was conducted by the use of the test data shown in Table I to determine the usefulness of the merged training data selected with the aid of the global correlation technique. Although 19 of the test talkers were represented in the merged training data, different samples (from training) were used for

Speaker	Word 0	1	2	3	4	5	6	7	8	9	10	11	12	13
1	-13	15	08	328	-76	-06	-32	-10	87	-26	-34	1[illegible]	-[illegible]	[illegible]
2	-13	48	16	180	-31	42	08	01	57	24	15	1[illegible]	[illegible]	[illegible]
3	-13	37	47	160	-72	21	-25	32	81	25	30	87	15	[illegible]
4	-15	-41	08	171	-53	-20	-3[illegible]	-26	76	-43	-31	72	-[illegible]	[illegible]
5	27	-30	43	175	-16	-17	01	04	103	25	-02	5[illegible]	04	6[illegible]
6	-09	-15	81	213	-30	16	-46	04	127	38	12	95	28	117
7	-20	00	63	185	-33	15	-4[illegible]	-27	117	-09	-1[illegible]	71	10	[illegible]
8	20	-54	05	117	-37	-22	37	20	106	-56	10	1[illegible]5	20	0[illegible]
9	-16	-53	32	91	-27	-14	-20	-22	129	64	31	6[illegible]	-01	3[illegible]
10	00	17	62	100	-82	-64	29	-15	87	53	37	12[illegible]	3[illegible]	[illegible]
11	-11	-33	97	212	-30	67	-73	09	100	05	05	48	48	53
12	-40	-23	32	176	-67	[illegible]	[illegible]	-[illegible]	87	-12	-14	87	1[illegible]	[illegible]
13	06	-12	55	167	-62	02	02	-[illegible]1	102	25	36	75	10	76
14	-30	-20	30	168	-52	29	-41	-09	75	-01	-05	47	29	22
15	-21	-27	77	170	-58	-42	66	-20	103	-01	-08	62	-59	81

COPY AVAILABLE TO DDC DOES NOT PERMIT FULLY LEGIBLE PRODUCTION

Figure 7. Correlation score matrix resulting from correlation of training data representing each word of each of 15 speakers with word 3 (the digit "three") of speaker 1.

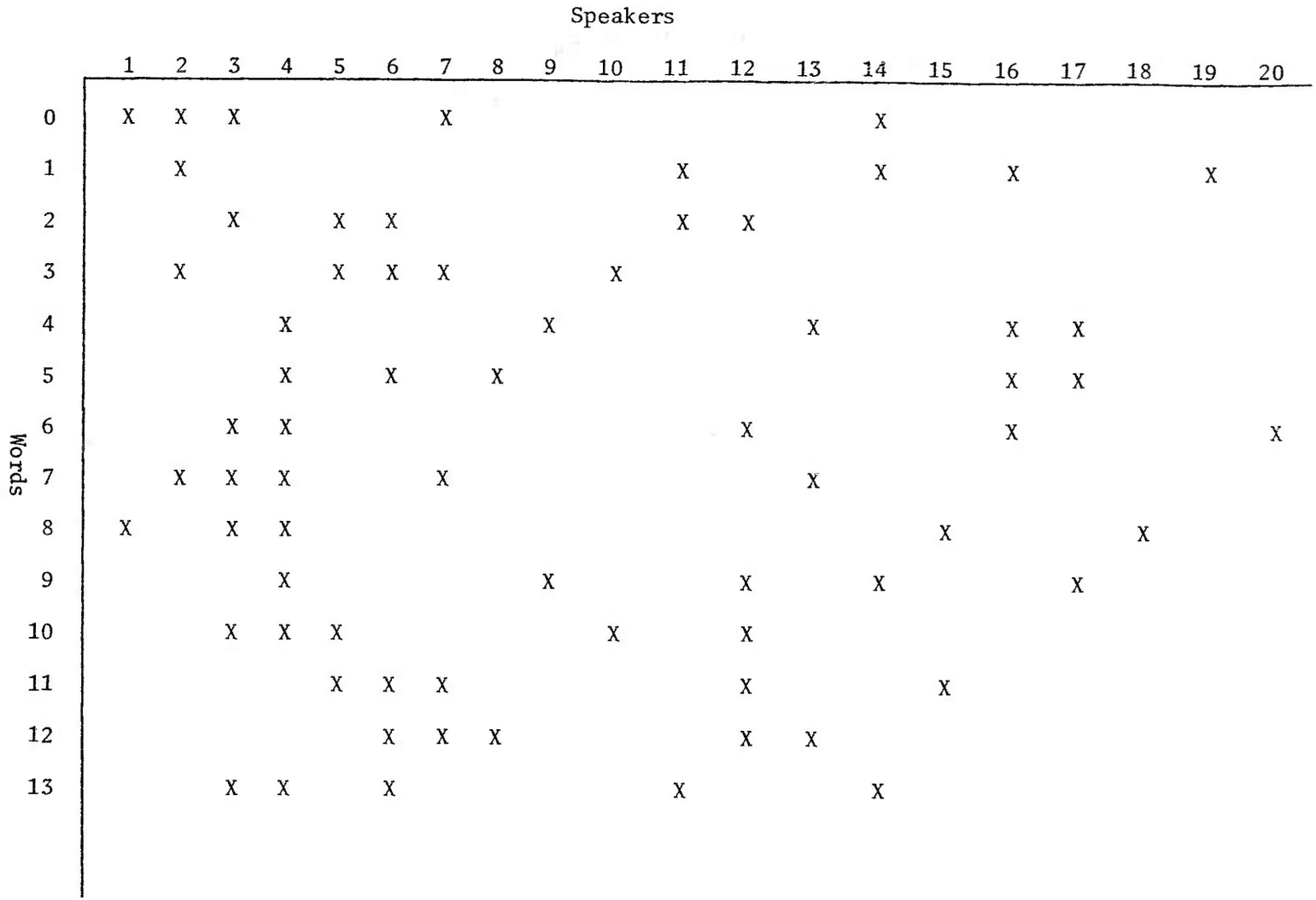


Figure 8. Composition of merged training samples used for five-sample merge.

testing. Recognition accuracy for 29 talkers was 97.15 percent overall. Accuracy for the 19 talkers who were represented in the training set was 97.45 percent. For the ten who were not represented in the training set, accuracy was 96.6 percent.

Next, a 70-word reference array set was constructed by the use of five entries for each vocabulary word. These five entries per word were from the same speakers whose training arrays were merged to form the 14 word reference array set for the test just described. A test of 29 talkers was then conducted with this 70 word reference array set. Results of this test indicate that the use of the separate entries taken from the highest ranking talkers is preferable to merging these samples. However, merging training data from a large set of talkers subsequently was found to be as effective for establishing reference arrays as choosing the best talkers from the set by the use of correlation matrices. This conclusion was derived from the results of another 29 talker test in which the reference arrays resulted from merging the complete 20 talker training data set. Results from the previously reported tests involving reference data from a merge of five training sets is shown below together with the results of the two new tests.

	Five Merged (14 Ref. Words)	Five Separate (70 Ref. Words)	20 Merged (14 Ref. Words)
20 talkers (in training set)	97.45	98.47	98.35
Nine talkers (not in training set	96.6	97.65	97.73
29 Talkers overall	97.15	98.22	98.16

Because the results achieved with the 20 talker merge were the best, all further tests during the program were accomplished by the use of 20 talker merge techniques.

E. One Word Training Sample Experiments

As outlined before, the use of single word training samples for the complete VICI vocabulary is precluded on the basis of the total time required for data input. However, a minimal training period of five seconds has been allowed in the operation of the advanced development model VICI. Therefore, it is possible to replace the reference array developed for a universal talker set with a single training sample each for a maximum of three digits by utilizing the five second training period. Several experiments were conducted to determine quantitatively the improvement in recognition accuracy afforded by the use of single training samples to augment a universal reference array. The most extensive of these experiments involved 50 speakers each inputting to VICI, 50 groups of four digits without any training data and then with a single repetition train on the digits one, three and nine.

The 50 talkers were recorded on audio tape and the tests were run from tape, subsequently. The test procedure did not exactly duplicate operational conditions insofar as the use of training digits was concerned. The tapes were recorded by each speaker with 50 four-digit groups spoken first, with single digits from zero through nine spoken as five sets following the digit groups. Table II is a list of the digit groups. In the tests with the three training digits, for each speaker the VICI system was first trained with a single sample for each of three digits taken from the sets of digits zero through nine. The 50 digit groups then followed. Therefore, as far as the speakers were concerned the training digits did not directly precede the code digits as they would in an operational situation. It can be reasonably inferred that this test procedure would result in accuracy slightly inferior to that realizable with a live input consisting of three training digits directly followed by a four digit code. Test results appear to bear out such an inference. Although in most cases recognition improved with the use of the three training digits a few speakers suffered slightly lower accuracy with the training digits. Results of these tests are shown in Table III. This table shows individual digit errors and corrections. In most cases there was only one digit error per group of four so that group error totals are only slightly lower than digit error totals. Any group of four digits in which any digit was mispronounced or garbled was not counted at all, i.e., all other digits in that group were ignored. Of a possible total of 2500 groups from 50 talkers, 2458 were usable groups. Recognition accuracy on a group basis went from 95.3 percent without training digits to 97.96 percent with three training digits. Individual digit accuracy of usable groups went from 98.77 percent without to 99.92 with training digits. Figures 9 and 10 are error matrices for these tests. Note that in order to show more clearly the error distribution, the correct responses have been omitted. Figure 9 shows the error matrix for the digit errors involved the digits 1, 3 and 9. Therefore, the single repetition train experiments with these three digits as the training digits could be expected to show significantly reduced errors. The results shown in Figure 9 prove out this assumption.

F. Recognition Networks for Universal Speaker Sets

The phoneme-like feature recognition networks included in the final VICI feature array previously discussed can be described by the use of logic equation as shown in Table IV. These logic equations can be translated into equivalent logic diagrams. The notational rules for these logic equations are as follows:

1. An expression of the form $(\int_{T1} XQ1 - \int_{T2} YQ2)$ indicates that the excitatory quantity Q_1 and the inhibitory (subtractive) quantity Q_2 are integrated with time constants T_1 and T_2 and employ gain factors X and Y, respectively.

2. The analytical expression for the binary AND function will be of the form C = A B, where C represents the digital output of the AND gate for the two inputs A and B which can be in analog or digital form.

3. The expression for a logical OR function will be the form C = A + B.

TABLE II LIST OF 50 FOUR DIGIT GROUPS USED FOR TESTS WITH TRAINING DIGITS

5 2 5 1	6 3 1 4	0 3 3 8
7 5 9 0	3 4 9 5	4 7 7 0
1 0 1 7	5 6 5 9	6 8 0 3
6 2 6 2	1 1 3 4	3 0 6 8
2 0 2 7	4 6 0 7	9 1 5 4
7 2 7 0	8 9 2 6	7 8 2 3
3 6 6 8	9 6 4 2	2 4 8 3
0 4 4 1	0 7 6 9	8 8 7 9
8 4 3 2	2 2 8 3	9 3 9 8
4 1 8 5	7 3 7 0	6 9 2 4
9 9 0 7	0 0 5 7	
5 8 3 1	1 4 0 6	
1 7 1 4	8 1 9 5	
0 9 8 6	9 7 4 6	
2 3 2 9	3 5 7 8	
6 7 0 9	2 1 2 5	
8 5 4 1	5 5 1 4	
3 8 6 2	1 6 1 5	
7 9 5 0	4 5 3 1	
4 2 9 6	5 0 8 3	

TABLE III RESULTS OF SINGLE DIGIT TRAIN EXPERIMENTS

Number of Speakers	Net Digit Errors Before Train	Net Digit Errors After Train	Errors Deleted With Train	Errors Added With Train
12	0	0	0	0
5	0	5	0	5
3	9	9	0	0
14	57	4	53	0
7	38	16	30	8
8	16	16	10	10
1	1	2	1	2
50	121	52	94	25

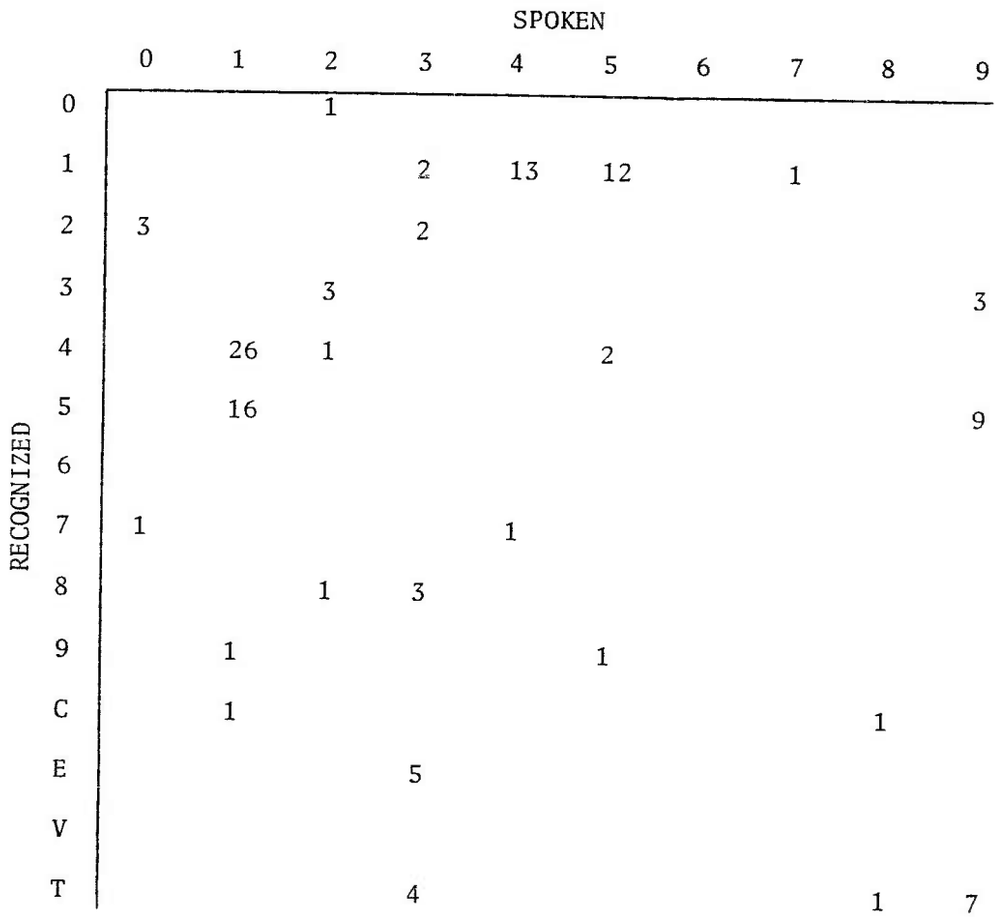


Figure 9 Error Matrix for 50 Speakers
Speaking 50 Groups of Four Digits Each (No Training)

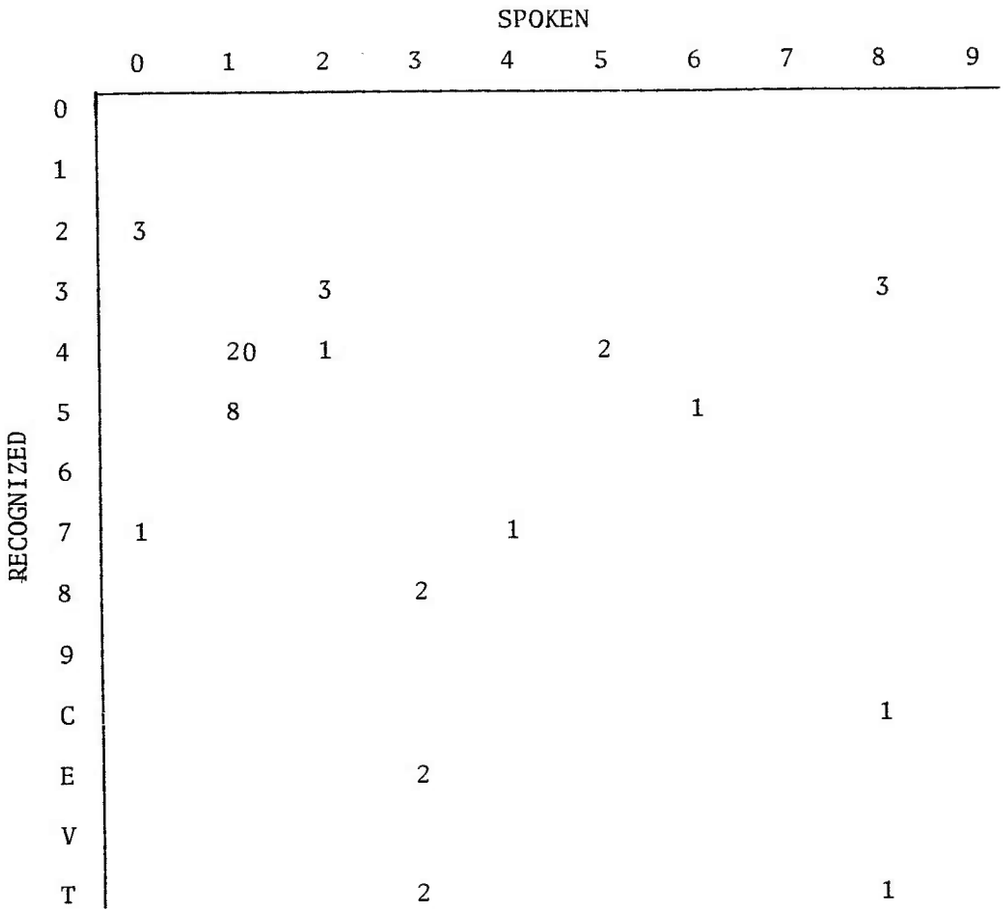


Figure 10. Error Matrix for 50 Speakers Speaking 50 Groups of Four Digits Each With Single Repetition Training on the Digits 1, 3, and 9

TABLE IV PHONEME-LIKE FEATURE RECOGNITION LOGIC EQUATIONS (SHEET 1 OF 4)

$$\text{V/VL} = \left[\left(\int_{5}^{3} \sum_{1} 1.7E - \int_{5}^{5} \sum_{4} 1.16E \right) + \left(\int_{5}^{3} \sum_{1} 1.7E - \int_{5}^{12} \sum_{10} 1.7E \right) \right] \cdot 1.27 \left[\left(\int_{5}^{4} \sum_{1} 1.7E - \int_{5}^{19} \sum_{15} 1.7E \right) + \left(\int_{5}^{11} \sum_{7} 1.7E - \int_{5}^{19} \sum_{16} 1.7E \right) - \int_{9} \left(\int_{5}^{19} \sum_{15} 1.4E - \int_{5}^{14} \sum_{10} 1.4E \right) \right] \cdot 1.27 \left[\left(\int_{5}^{14} \sum_{11} 1.7E - \int_{5}^{18} \sum_{15} 1.7E \right) + \left(\int_{5}^{9} \sum_{6} 1.7E - \int_{5}^{13} \sum_{10} 1.7 \right) + \left(\int_{5}^{4} \sum_{1} 1.7E - \int_{5}^{11} \sum_{8} 1.7E \right) - \int_{9} \left(\int_{5}^{19} \sum_{15} 1.4E - \int_{5}^{14} \sum_{10} 1.4E \right) \right]$$

$$\text{BV} = \left[\left(\int_{5}^{8} \sum_{6} 2.75E - \int_{5}^{11} \sum_{9} 2.75E \right) + \left(\int_{5}^{9} \sum_{7} 2.75 - \int_{5}^{12} \sum_{10} 2.75E \right) + \text{NSB8} \right]$$

$$\text{i} = \left[\left(\int_{5}^{14} \sum_{9} 1.7E - \int_{5} 1.7 \sum_{2}^{3} E \right) \cdot \text{NSB2} \cdot \text{NSB3} \cdot \overline{(\text{MAX5+6+7})} \cdot \text{V/VL} \right] \cdot \left[\underline{\text{BRST}} \right] \cdot \left[(\text{PSB7} + 10) \cdot \text{PSB8} \cdot \text{PSB9} \cdot \left(\sum_{3}^{4} \text{NSB} - \sum_{6}^{7} \text{NSB} \right) \cdot \left(\sum_{8}^{10} \text{PSB} - \text{NSB4} - \text{NSB7} \right) \cdot \left(\int_{5}^{14} \sum_{11} 2.75E - \int_{5}^{19} \sum_{17} 2.75E \right) + \text{NSB4} \cdot \left(\int_{5}^{10} \sum_{7} 1.75E - \sum_{15}^{17} 1.75E \right) \cdot 1.16 \left(\sum_{9}^{10} \text{PSB} - \sum_{7}^{8} \text{NSB} - \sum_{7}^{8} \text{PSB} \right) + 1.16 \left(\sum_{7}^{9} \text{PSB} - \sum_{10}^{12} \text{PSB} - \sum_{7}^{8} \text{NSB} \right) \right]$$

TABLE IV PHONEME-LIKE FEATURE RECOGNITION LOGIC EQUATIONS (SHEET 2 OF 4)

$$I = V/VL \cdot NSB3 \cdot NSB4 \cdot PSB7 \cdot PSB8 \cdot (1.15\ PSB1 - 1.7\ PSB2) \cdot \left(\int_{5}^{} \sum_{10}^{13} 2.75E - \int_{5}^{} \sum_{16}^{19} 2.75E\right) \cdot$$

$$\left(\int_{5}^{} \sum_{9}^{11} 2.75E - \int_{5}^{} \sum_{6}^{8} 2.75\ E\right)$$

$$\varepsilon_1 = V/VL \cdot PSB\ (7+8) \cdot NSB4 \cdot (1.16\ PSB1 - 1.7\ PSB3) \cdot \left(\int_{5}^{} \sum_{9}^{11} 5.63E - \int_{5}^{} \sum_{5}^{7} 5.63E\right)$$

$$\varepsilon_2 = V/VL \cdot MAX3 \cdot (MAX9 + MAX10) \cdot NSB4$$

$$\text{ɜ} = V/VL \cdot BV \cdot NSB9 \cdot NSB10 \cdot NSB11 \cdot (NSB8 + NSB12) \cdot \left(1.4 \sum_{8}^{9} NSB - 1.4 \sum_{6}^{7} NSB\right)$$

$$r = NSB3 \cdot NSB9 \cdot NSB10 \cdot NSB11 \cdot (NSB8 + NSB12) \cdot \left(1.4 \sum_{8}^{9} NSB - 1.4 \sum_{6}^{7} NSB\right) \cdot$$

$$\left(1.4 \sum_{3}^{4} NSB - 1.4 \sum_{6}^{7} NSB\right) \cdot \left(\int_{5}^{} \sum_{7}^{9} 2.75E - \int_{5}^{} \sum_{10}^{12} 2.75E\right) \cdot \left(1.15 \sum_{8}^{10} NSB - 1.15 \sum_{2}^{4} NSB\right)$$

TABLE IV PHONEME-LIKE FEATURE RECOGNITION LOGIC EQUATIONS (SHEET 3 OF 4)

$$W = V/VL \cdot BV \cdot \left[\left(\int_{5}^{3} \sum_{1} 1.15E - \int_{5}^{10} \sum_{7} 1.15E \right) \cdot \left(\int_{5}^{6} \sum_{4} 1.15\ NSB - \int_{5}^{11} \sum_{8} 1.15\ NSB \right) + \left(\int_{5}^{3} \sum_{1} 1.15E - \int_{5}^{9} \sum_{5} 1.15E \right) \cdot \left(\int_{5}^{5} \sum_{3} 1.15\ NSB - \int_{5}^{11} \sum_{8} 1.15\ NSB \right) \right]$$

$$UVNLC = \overline{V/VL} \left[\left(\int_{14}^{19} \sum_{15} 1.4E - \int_{14}^{14} \sum_{10} 1.4E \right) + \left(\int_{14}^{19} \sum_{15} 1.4E - \int_{14}^{9} \sum_{5} 1.4E \right) \right] \cdot \left[\underline{\quad BRST \quad} \right] \cdot \left[\left(\int_{14}^{19} \sum_{15} 1.4E - \int_{14}^{4} \sum_{1} 1.7E \right) \right]$$

$$S = UVNLC \left[PSB\ (5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13) + PSB\ (6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14) + PSB\ (7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 14 \cdot 15) \right]$$

$$\Lambda = V/VL \cdot BV \cdot \left[\left(1.4 \sum_{6}^{7} PSB - 1.4 \sum_{8}^{9} PSB \right) + \left(1.4 \sum_{8}^{9} NSB - 1.4 \sum_{6}^{7} NSB \right) \right] \cdot \left[PSB\ (2+3) + PSB\ (3+4) \cdot PSB2 \cdot \left(\int_{5}^{13} \sum_{10} 2.75E - \sum_{16}^{19} 2.75E \right) \right]$$

TABLE IV PHONEME-LIKE FEATURE RECOGNITION LOGIC EQUATIONS (SHEET 4 OF 4)

$$n = \left[\text{NSB}\ (1{\cdot}2{\cdot}3) \cdot \text{EG}_1 \cdot \text{EG}_2\right] + \left[\text{EG}_1 \cdot \text{V/VL} \cdot \text{NSB}\ (2{\cdot}3{\cdot}4) \cdot \underline{\text{BRST}} \cdot \left(\int_{5}^{4}\sum_{1} 1.7\text{E} - \int_{5}^{9}\sum_{6} 1.7\text{E}\right) \cdot \left(\int_{2.2}^{4}\sum_{1} 1.15\ \text{NSB} - \int_{2.2}^{4}\sum_{1} 1.15\ \text{NSB} - \int_{2.2}^{9}\sum_{6} 1.15\ \text{NSB} - \int_{2.2}^{9}\sum_{8} 1.15\ \text{PSB}\right)\right]$$

$$\text{Energy Gap} = \text{EGap}_1 + \text{EGap}_2 = \left[\int_{2.2}^{3.4} \int_{55} \left(\int_{5}^{9}\sum_{5} .25\text{E} - \underline{5.63\ \text{UVNLC}}\right) - \int_{2.2}^{3.4} \int_{5}^{9}\sum_{5} .25\text{E}\right] + \left[\left(\int_{2.2}^{3.4} \int_{35} \left(\int_{5}^{14}\sum_{10} .25\text{E} - \underline{5.63\ \text{UVNLC}}\right) - \int_{2.2}^{3.4} \int_{5}^{14}\sum_{10} .25\text{E}\right)\right]$$

$$\underline{\text{Slope Gap}} = \int_{2.2}^{2.3} \int_{35} (x) - \int_{2.2}^{3.4} (x)$$

$$\text{where } (x) = \left[\sum_{7}^{10} 0.25\ \text{NSB} - \int_{19} (0.5\ \text{NSB1} + 0.4\ \text{NSB2} + 0.25\ \text{NSB3} + 0.4\ \text{NSB4} + 0.5\ \text{PSB5})\right] \cdot \left[\text{BRST}\right]$$

4. The summation symbol $\sum_{m}^{n} Q$ will be used to indicate a plurality of (analog) input signals of the same type to an ATL element. In each case Q represents the type of input signal, m and n represent the interval over which the feature is summed.

5. The networks or portions of networks which were constructed or modified expressly for the VICI vocabulary are underlined with broken lines.

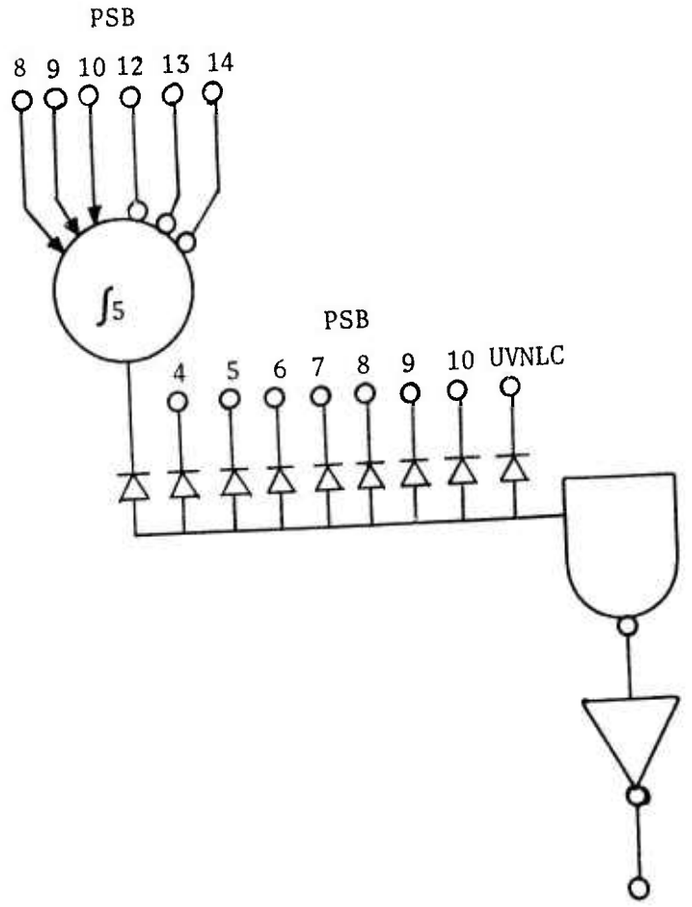
An example of the relationship between the logic diagram and the logic equation for a particular feature recognition network is shown in Fig. 11. The network shown in the figure was designed to recognize /ʃ/ in a conventional VIP-100 preprocessor. This phoneme is not currently included in the VICI feature array because it does not appear in the VICI vocabulary. The network includes as inputs both binary and analog representation of positive slopes. Design considerations for this network are explained in the following paragraph.

This fricative consonant is characterized in wide-band speech by broad noise-like frequency bands above 1k Hz with a broad energy peak in the 2-3 kHz region. The resultant primary features useful for the detection of this characteristic are positive slopes (PSB) up through channels 10 or 11 of the VIP preprocessor. This phoneme is separated from the similar fricative, /s/ by the strength of positive slopes in channels 8 through 10 as compared with the positive slopes in channels 12 through 14. The phoneme /ʃ/ with a lower frequency concentration of energy in its spectrum as compared with /s/ can be expected to have stronger slopes in channels 8 thorugh 10 as compared with channels 12 through 14. This separation is accomplished by the ATL element in the network as shown in the figure. The integration time constant associated with the inputs of the ATL element is 5 milliseconds. An input resistor value of 42.2K ohms is used for each input resulting in a gain factor of 0.8 times for each input. The unity gain input resistance for an ATL element is 34K ohms. Lower values of input resistance will therefore result in gains of greater than one. Binary representations of positive slopes in channels 4 through 10 together with the unvoiced noise-like (UVNLC) feature typical of fricatives are ANDed together with the output of the ATL element.

Networks for the Zero Crossing feature included in the VICI feature array and another special feature, labeled BRST in several equations are not described in Table IV because they are not easily expressable in logic equations. The BRST network is a specialized arrangement of digital logic designed to minimize the deleterious effects of a burst at the end of the digit 8 or the control word TERMINATE. These bursts are quite variable with regard to spectrum and frequency of occurrence from talker to talker. Therefore, a major effort was successfully made to ignore the burst when it occurred.

G. VICI Software

The software package developed for use with the VICI system was based upon a standard VIP-100 program for recognizing a vocabulary of up to 76 words with storage capacity for training data for one speaker at a time in the 8K of core memory supplied with the system. Included as integral parts of this program were the training algorithm, recognition algorithm and an output routine



$$\int = \text{UVNLC} \cdot \text{PSB4} \cdot \text{PSB5} \cdot \text{PSB6} \cdot \text{PSB7} \cdot \text{PSB8} \cdot \text{PSB9} \cdot \left(\int_{5}^{10} \sum_{8} \text{PSB} - \int_{5}^{14} \sum_{12} \text{PSB}\right)$$

Figure 11. Logic diagram and equivalent logic equation for /ʃ/ recognition net word.

for driving a 16 character Burroughs Self-Scan display and an ASR 33 Teletype. This large vocabulary capacity allowed the experiments with the use of up to five alternate reference arrays for each of the 14 words in the VICI vocabulary, thus requiring in effect a 70 word vocabulary. Subsequent to these experiments with multiple reference arrays, merging of reference arrays of large number of talkers was found to provide equal recognition accuracy for large number of speakers. The use of a small reference array set provides faster inputting of data. The approximate recognition time of the VIP-100 software for an input word is 100 ms with a vocabulary size of 14 and 300 ms with a vocabulary size of 70. This time is processing time required by the minicomputer after the cessation of the word and does not include the actual time required to pronounce the word. The computer can accept a new input word during the decision processing for a previous input. However, the larger size vocabulary requires significantly greater processing time because of the necessity in the recognition process of correlating the feature array of the input word with each reference array entry. Each correlation requires approximately 4 ms.

A major modification of the conventional VIP-100 recognition software was made to allow certain additional correlations to take place after the initial recognition decision. These additional correlations, known as "second-look", involve only the initial portion of the feature array of an input word and selected reference arrays. To facilitate an explanation of this special correlation, a review of the normal VIP-100 correlation process is in order.

After the input word is time normalized into 16 time slots the resultant array which is 32 features wide (composed of digital ones and zeros) is compared with each stored reference array. Similarities and dissimalarities in each array are compared and appropriately weighted and the net result provides a weighted correlation product. Two other correlation products are produced for each reference array after shifting the input array $\pm$ one time slot. The highest correlation product for each stored reference array is then compared with the highest products for each other word. The overall maximum product decides which word is recognized.

The second-look correlation routine, when made operative, correlates in a similar manner the first five time slots only of the input array and a selected reference array. The highest correlation product of this special correlation is then multiplied by four and is added to the original product for the particular reference array selected for the special correlation. This special routine is used for selected words only. Second-look takes place only if the initial correlation routine selects the input word as zero, one, five, eight, or TERMINATE. For each of these words a different set of reference arrays are involved. If the initial correlation choice is the digit zero, the input array is recorrelated against the reference array for zero and for two, because most confusions involving the digit zero have been with the digit two. Likewise, if a one is recognized the second look correlation occurs with the reference arrays for one, four and five. If a five is recognized second-look occurs for five and nine. If eight is chosen the reference arrays for eight, two, and three are recorrelated. In a similar manner, TERMINATE initiates second-look for TERMINATE and three. The second-look routine has been found to be quite effective in increasing accuracy for some talkers.

Figures 12 and 13 illustrate by means of error matrices the extent of such improvement. Figure 12 is the error matrix of a test without second-look of 34 speakers reading the random word list shown in Table I. Figure 13 is a similar test of the same speakers and words with the second-look configuration as shown above except for the TERMINATE-three combination which was not added until this test was completed. Also a recognition logic change was made which effected principally the control word TERMINATE. This change resulted in fewer misrecognitions of TERMINATE in the second test but more instances of the digit three being misrecognized as TERMINATE. The subsequent addition of the TERMINATE-three combination to the second-look routine virtually eliminated this confusion as reference back to Figure 9 will disclose. This figure is the error matrix for 50 speakers resulting from digit groups.

The second-look routine was found to be especially helpful in reducing the number of times the digit four was misrecognized as one (21 times without second-look, 3 times with). The number of 3-8 confusion and 9-5 confusion was also reduced as is illustrated by comparison of Figures 12 and 13.

RECOGNIZED \ SPOKEN	0	1	2	3	4	5	6	7	8	9	C	E	V	T
0														
1				1	21	3		1						
2										1				6
3			7							1		1		5
4		12					1							
5		6		1				1		22				6
6												1		
7													1	1
8			11	13							2	1		
9								1						1
C							1					5	3	
E									2		5			
V						1								
T				2										
Rej.							1		2			1		

Figure 12 Error Matrix for 34 Speakers
Reading VICI Vocabulary List Without Second-Look

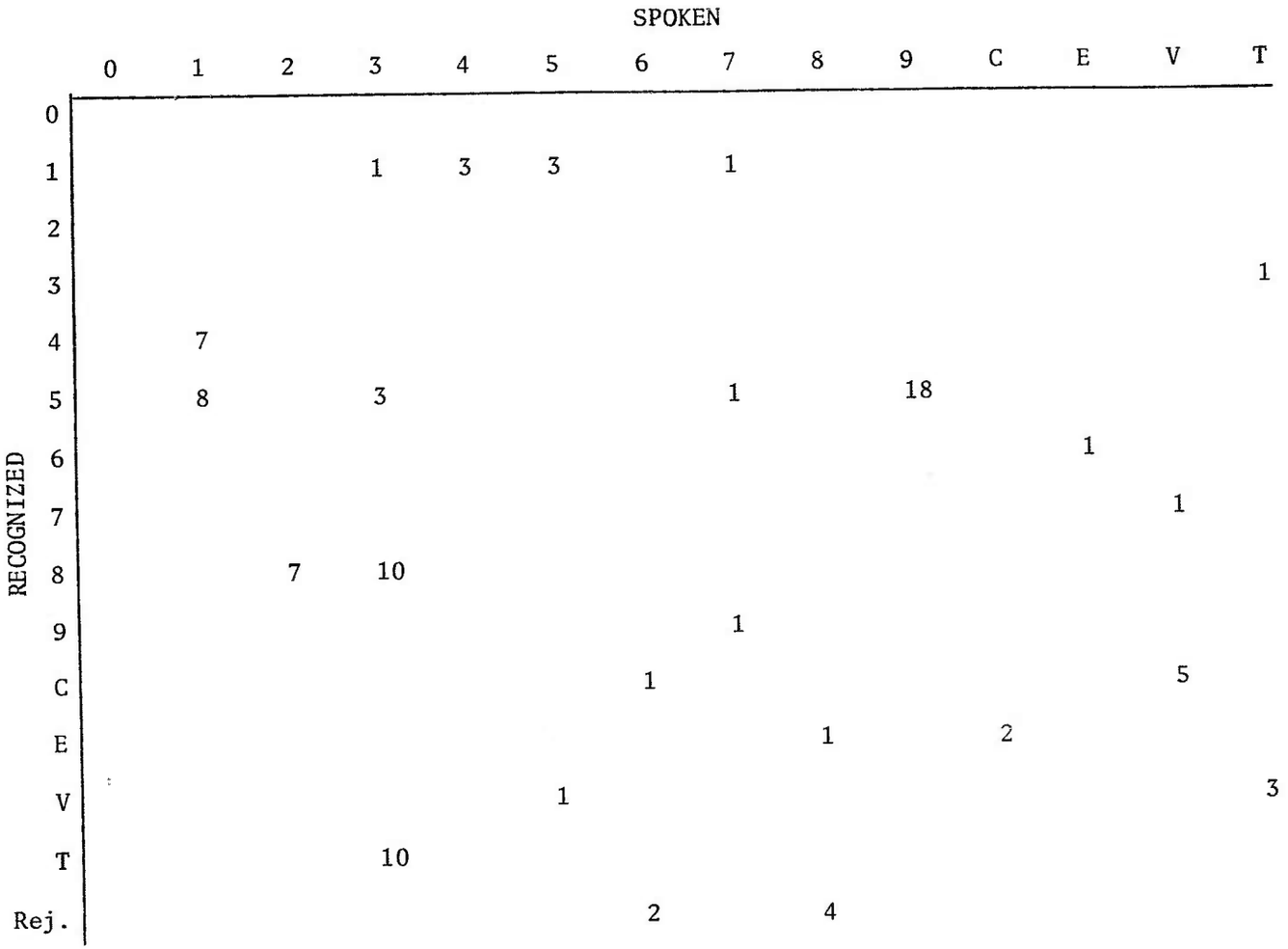


Figure 13 Error Matrix for 34 Speakers
Reading VICI Vocabulary List with Second-Look

Section III

FINAL SYSTEM TESTS

A. Background of Test Data

Final testing of the VICI system to establish performance levels was conducted by the use of both tape recorded and live inputs from a total of 83 male talkers ranging in age from 16 years to 65 years. Tape recordings were made of digits and control words spoken by 65 talkers over a period of eight months from August 1974 to March 1975. In addition, special training data recordings of 20 talkers, all TTI employees, were made in July 1974. This training data from which universal reference arrays were derived consisted of 10 repetitions of each VICI vocabulary word as spoken by each of the 20 speakers. These same 20 talkers also later recorded independent test data. No less than two weeks elapsed between the time any test and any training data were recorded by the same talker. All recordings were made with Telex model 1200 noise cancelling microphones. Figure 14 is a frequency response plot of one of the two microphones used for these recordings. The other microphone had a similar response.

The test data initially recorded were of the list of digits and control words in random order as shown in Table I. Each speaker recording this list read it two times, thus producing a total of 280 words. A total of 41 talkers recorded this list including the 20 who had previously recorded the original training data. Data recorded by four Air Force employees of Wright-Patterson Air Force Base in November 1973 for another contract were also used in tests involving the Table 1 list. These four Air Force employees spoke the digits and two control words, ERASE and TERMINATE, 10 times per word in random order. These latter four recordings were also made with a Telex 1200 microphone. The list of 50 four-digit groups shown in Table II was recorded by 50 speakers, including 18 of the 20 who recorded original training data and 30 of the 41 who recorded the list of random digits and control words shown in Table I. For the live tests, the four-digit group list was expanded to 75 groups as shown in Table IV. Two live tests were conducted, one at TTI just prior to delivery of the VICI system to RADC and one at RADC upon delivery. The first live test was a 10 talker test. Nine of these 10 talkers were TTI employees, all of whom had participated in recording original training data, in the recording of the list of digits and control words in random order and in the recording of the 50 four-digit groups. The tenth participant in the live test at TTI was RV, an RADC representative, who supervised the test. The second live test, held at RADC included 21 speakers, 20 of whom were civilian employees of RADC and military personnel stationed at RADC. The twenty-first speaker in the test at RADC was PS, the TTI project engineer for VICI who also participated in the live test at TTI.

B. Final Testing From Tape

The final tests with tape recorded test data included two groups of data, the list of digits and control words in random order, and the four-digit groups. The final test with the former data culminated a series of tests with this data base during the development of the system. The four-digit test conducted sub-

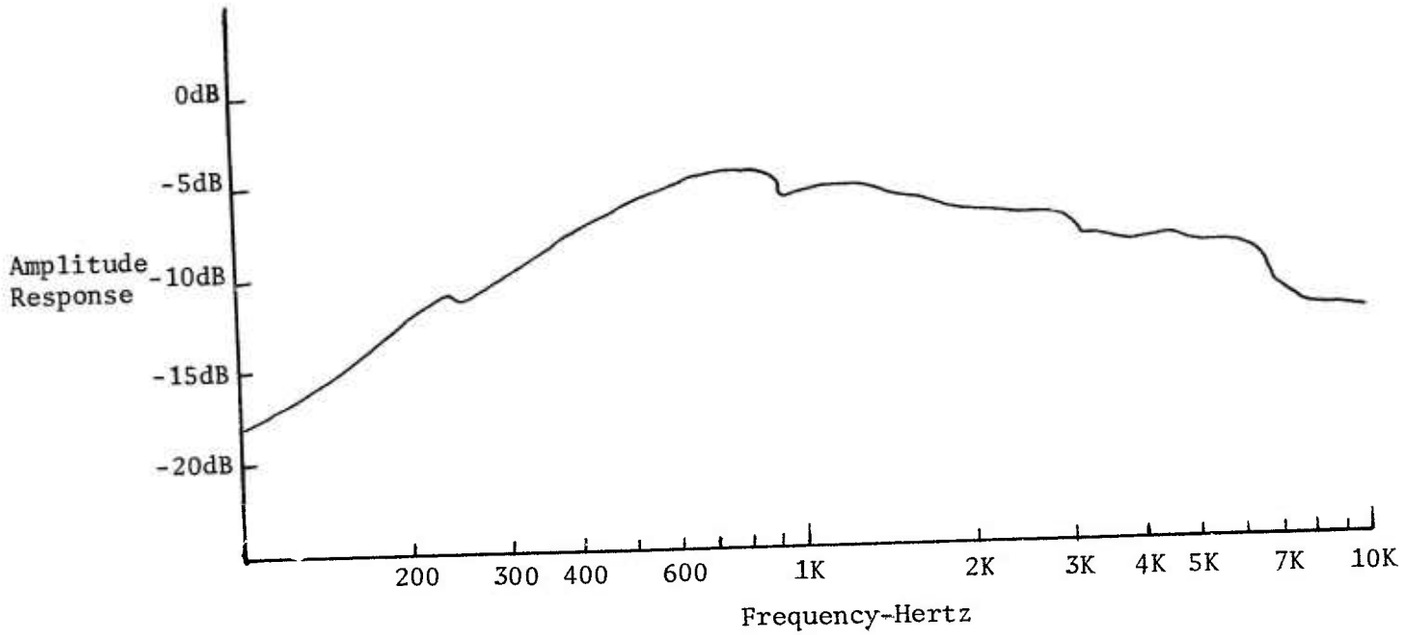


Figure 14 Measured near-field frequency response of Telex Model 1200 microphone used for making VICI data tape recording. Microphone measured at 1/4" distance from output orifice of a calibrated Plane-Wave-Tube.

TABLE V LIST OF 75 FOUR-DIGIT GROUPS USED FOR FINAL LIVE TESTS AT TTI AND AT RADC

5 2 5 1	6 3 1 4	0 3 3 8
7 5 9 0	3 4 9 5	4 7 7 0
1 0 1 7	5 6 5 9	6 8 0 3
6 2 6 2	1 1 3 4	3 0 6 8
2 0 2 7	4 6 0 7	9 1 5 4
7 2 7 0	8 9 2 6	7 8 2 3
3 6 6 8	9 6 4 2	2 4 8 3
0 4 4 1	0 7 6 9	8 8 7 9
8 4 3 2	2 2 8 3	9 3 9 8
4 1 8 5	7 3 7 0	6 9 4 2
9 9 0 7	0 0 5 7	1 5 2 5
5 8 3 1	1 4 0 6	0 9 5 7
1 7 1 4	8 1 9 5	7 1 0 1
0 9 8 6	9 7 4 6	2 6 2 6
2 3 2 9	3 5 7 8	7 2 0 2
6 7 0 9	2 1 2 5	
8 5 4 1	5 5 1 4	
3 8 6 2	1 6 1 5	
7 9 5 0	4 5 3 1	
4 2 9 6	5 0 8 3	
2 4 9 6	0 7 3 7	
8 9 3 9	3 8 2 2	
9 7 8 8	9 6 7 0	
3 8 4 2	2 4 6 9	
3 2 8 7	6 2 9 8	
4 5 1 9	7 0 6 4	
8 6 0 3	4 3 1 1	
3 0 8 6	9 5 6 5	
0 7 7 4	5 9 4 3	
8 3 3 0	4 1 3 6	

sequently simulated to an extent operational conditions and also tested the use of single training samples as outlined in Section II, E. The four-digit group test did not exactly simulate operational conditions because there was no way to verify and correct errors.

1. Random Digit and Control Word Test Results

The final test of 280 words, digits and control words in random order was conducted shortly before the four-digit group tests in order to determine overall word accuracy without the aid of one word training samples. Overall word accuracy for 45 speakers was 98.83 percent. The control word recognition accuracy in this test was 99.2 percent with the digit recognition accuracy slightly lower at 98.7 percent. Figure 15 is the error matrix resulting from this test.

2. Four-Digit Group Tests

A test of 50 speakers each uttering 50 groups of four digits each was conducted to simulate operating conditions. Complete results of this test are given in Section II, E which describes experiments with single word training samples. These experiments were conducted in conjunction with the final test of four-digit groups. To recapitulate these test results, without any training digits the single digit accuracy was 98.77 percent, virtually the same as in the test described above. The four-digit group accuracy was 95.3 percent. Because the tests were conducted from tapes, there was no error correction.

C. Final Testing with Live Inputs

The final test of the VICI system with live inputs was conducted in two parts as previously indicated, a 10 talker test at TTI before equipment delivery and a 21 talker test at RADC upon equipment delivery. These tests were conducted in order to determine whether the system could meet the required accuracy of 98 percent for four-digit code groups with a combined 2 percent rejection and error rate. Error correction by the use of a visual display was incorporated in these tests. This display was adjusted to show up to and including four digits at a time. With error correction for both tests, all speakers were able to successfully input all of the 75 digit groups shown in Table IV for an accuracy of 100 percent. Results of the two tests are shown in Tables V and VI. Encoding time durations were not recorded for the tests at TTI but were recorded for the RADC test. The average encoding time of 6.24 seconds in the latter test was well within the allowed 10 second period.

In each test two of the four control words were used to perform error correction. The word ERASE caused the last digit entry to be deleted from the display while CANCEL deleted all previous entries. For the test at RADC the word VERIFY caused the display screen to be blanked and the word "IN" appeared indicating to observers of the test that the speaker had verified the correctness of the four digit entry. In an operational application, the verified four-digit group could be outputted to the BISS system at this time by the same control word. Each speaker was instructed to say the control word VERIFY after he was satisfied that the four-digit group was entered correctly, thus

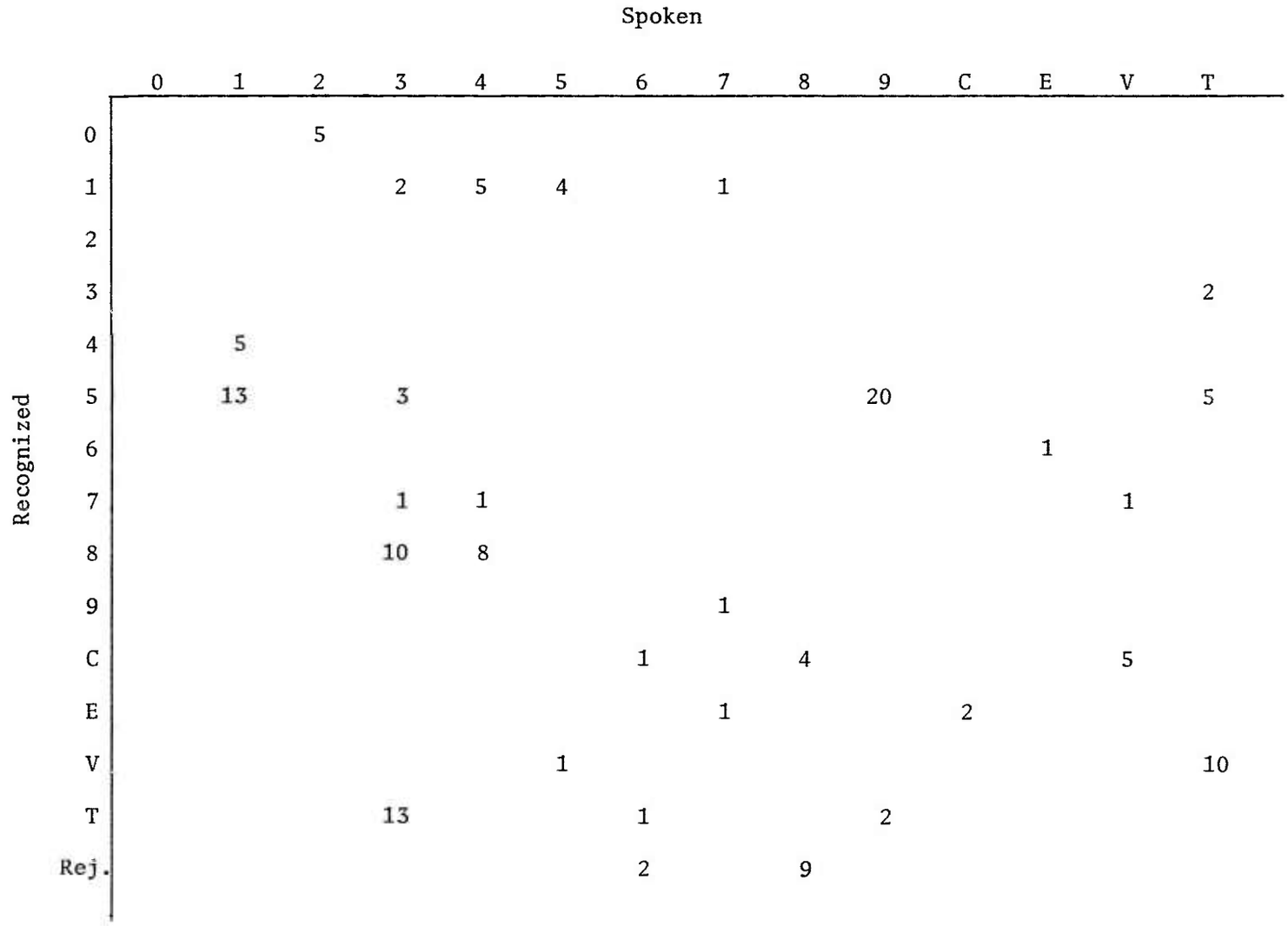


Figure 15 Error matrix of 45 speakers each uttering 280 digits and control words in a random arrangement.

TABLE VI RESULTS OF LIVE TEST HELD AT TTI WITH 10 SPEAKERS EACH INPUTTING 75 FOUR-DIGIT GROUPS

Speaker	Incorrect Digits	Digit Error Percent	Incorrect Groups	Group Error Percent
WL	21	7	18	24
EG	12	4	12	16
MP	2	0.67	2	2.67
PS	3	1	3	4
MH	2	0.67	2	2.67
LS	1	0.33	1	1.33
MW	3	1	3	4
AT	1	0.33	1	1.33
AP	4	1.33	4	5.33
RV	4	1.33	4	5.33
Total	53		50	
Average	5.3	1.76	5.0	6.8

TABLE VII RESULTS OF LIVE TEST HELD AT RADC WITH 21 SPEAKERS EACH SPEAKING 75 FOUR-DIGIT GROUPS

Speaker	Incorrect Digits	Digit Error Percent	Incorrect Groups	Group Error Percent	Encoding Times Over 10 sec.	Min. Encoding Time - sec.	Avg. Encoding Time - sec.
MK	3	1.00	3	4.00	1	4.0	6.18
TM	4	1.33	4	5.33	4	5.0	7.14
DJ	7	2.33	7	9.33	7	4.0	6.51
BR	1	.33	1	1.33	1	4.5	5.28
PS	0	0	0	0	0	2.8	3.43
RIJ	19	6.33	18	24.00	12	4.5	7.30
GSJ	0	0	0	0	0	5.0	5.83
RH	10	3.33	10	13.33	9	6.0	8.26
CC	10	3.33	10	13.33	10	4.5	6.75
TD	8	2.66	6	8.00	4	5.0	6.45
CM	0	0	0	0	0	5.0	6.55
KN	8	2.66	8	10.66	7	4.0	5.72
BB	1	.33	1	1.33	1	3.5	5.35
ES	21	7.00	20	26.66	11	3.5	7.06
JP	1	.33	1	1.33	1	3.5	4.66
BC	6	2.00	5	6.66	3	4.0	4.85
BS	7	2.33	7	9.33	5	3.5	5.67
MC	14	4.66	13	17.33	14	4.0	6.64
DC	1	.33	1	1.33	3	3.0	5.16
JS	3	1.00	3	4.00	4	6.0	7.37
JL	18	6.00	16	21.33	19	5.5	8.91
Total	142		134		116		
Average	6.76	2.25	6.38	8.50	5.52	4.32	6.24

effectively making each group a five word entry. The statistics in Table IV involving the time required for entering codes include the entry of the word VERIFY following the four-digit group. Tables V and VI indicate the total number of digit and group errors which occurred prior to verification and correction. With verification and correction all speakers were able to enter all codes correctly. The individual digit recognition accuracies of 98.24 percent for the test at TTI and 97.75 percent for the test at RADC were in good agreement with single digit accuracy from tape.

Section IV

CONCLUSIONS AND RECOMMENDATIONS

A. Conclusions

The VICI system is being developed as a front end for the BISS automatic speaker verification system to provide a reliable fully automatic means for entering speaker verification data. The VICI has demonstrated high accuracy capability as an isolated-word recognition system for the English digits plus four control words. Recognition accuracy for individual digits is approximately 98 percent without error correction or any individual adaptation for a population of 85 male talkers ranging from 16 to 65 years of age. The design requirement of 98 percent accuracy for the input of four-digit code groups with error correction by the speaker has been exceeded.

In two live speaker tests, everyone of a total of 30 speakers each uttering 75 code groups of four digits each was able, with the aid of error correction, to input correctly all of the code groups, for 100 percent accuracy. The average time required to speak and verify each code group and the word VERIFY was 6.24 seconds for 21 talkers in one of these tests. Time was not recorded for the other test. This average entry time included correction of misrecognized digits when necessary. Error correction was accomplished by allowing each speaker to view on a display the recognition decision immediately after it was spoken (within .1 to .2 seconds). A recognition error could then be corrected by saying the word ERASE which deleted the incorrect digit, and saying the digit again. Occasionally, speakers would pronounce several digits or a complete group before realizing an initial error. In this instance, the group could be deleted by saying CANCEL.

The performance levels achieved by the VICI system were made possible by a number of modifications to a basic VIP-100 speech recognition system manufactured by Threshold Technology Inc. for commercial applications. The modifications, in both hardware and software, were tailored to the limited vocabulary set required. Extensive experimentation was conducted to determine the optimum configuration of the reference array data to be used in the VICI system for recognition of the digits and control words. The merging of training arrays generated by several speakers to form a master reference array has been found to be quite useful. The final VICI reference array set resulted in a merge of training arrays from 20 speakers. This reference array set was then used for all final testing.

B. Recommendations

The VICI system as presently constituted operates with a headband-mounted noise-cancelling microphone connected by a high quality wire or radio link to the preprocessor. For field use with the BISS system, it is likely that operations under less than ideal conditions will be necessary. The operational constraints could include a handheld microphone or telephone-handset microphone with a 300 to 3 kHz wire link connecting the microphone at a remote entry point to the VICI system located at a central location. An investigation into the use of a handheld microphone transmitting speech over a wire-line could result

in modification of the VICI to allow operation under such conditions.

The VICI system has been developed for General American male speakers. Operational conditions will undoubtedly include inputs by females as well as males. The VIP-100 speaker-dependent word recognition system upon which VICI is based has shown excellent performance with female as well as male speakers in numerous commercial applications. Therefore, modifications of the VICI system to accept female as well as male speakers should not be difficult. A possible approach to fully universal speaker operation would be the use of two or more alternate reference array sets, at least one for each sex. The use of alternate reference arrays for recognition of large numbers of male talkers has been successfully tested and was described in Section II of this report. These techniques should be extended to allow male and female talkers to use the system.

The very high accuracy of the VICI system was enhanced by the use of manual error correction by the speakers testing the system live. An alternative approach to manual correction is the use of error correcting codes and/or check digits. A study of the use of such codes should be conducted in order to minimize the amount of manual error correction necessary and thereby speed the inputting of digits.

Certain digit confusions such as 1-4 and 3-8 were found to occur in the final testing with enough frequency to cause a few speakers some annoyance. Also, the use of the word "niner" for the digit nine which was done routinely in the testing has been deemed undesirable from an operational standpoint by RADC project personnel. Therefore, these problem areas should be given special attention in any program for improvement of the VICI system.

MISSION
of
Rome Air Development Center

RADC is the principal AFSC organization charged with planning and executing the USAF exploratory and advanced development programs for information sciences, intelligence, command, control and communications technology, products and services oriented to the needs of the USAF. Primary RADC mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, and electronic reliability, maintainability and compatibility. RADC has mission responsibility as assigned by AFSC for demonstration and acquisition of selected subsystems and systems in the intelligence, mapping, charting, command, control and communications areas.